Weed Mapping with Convolutional Neural Networks on High Resolution Whole-Field Images

Yuemin Wang[1] Thuan Ha[2] Kathryn Aldridge[2] Hema Duddu[2] Steve Shirtliffe[2] Ian Stavness[1]

[1]Department of Computer Science
University of Saskatchewan, Canada
yuw422@usask.ca, ian.stavness@usask.ca
[2]Department of Plant Sciences
University of Saskatchewan, Canada
thuan.ha@usask.ca, ksa083@usask.ca, hema.duddu@usask.ca, steve.shirtliffe@usask.ca

Abstract

Weed mapping is a technique used to identify and locate harmful weed plants in farm fields. Accurate weed mapping enables targeted herbicide application and helps plant scientists to estimate the effectiveness of field experiments. In this paper we discuss a highly practical and effective working pipeline to weed map a wheat field combining GIS and deep learning technology. This pipeline is an end-to-end process including using an unoccupied aerial vehicle (UAV) to collect ultra-high definition whole-field images, labelling and training deep learning models and an efficient evaluation process for the resulting weed map. We show that our method can generate accurate pixel-wise weed maps by only training on small regions of the field, and can generalize well when making predictions back on the larger whole-field orthomosaic image.

1. Introduction

Weed management remains one of the primary concerns for modern agronomy and uncontrolled weeds lead to yield loss [1] and reduced carbon sequestration [10] in crop fields. The common approach to weed control involves uniform broad spraying of herbicide on the field. Even though this method is effective, broad herbicide application creates a high cost for the farmers and the large quantity of chemicals used also has potential for negative impact on the environment. These problems led to increasing interest in precision spraying due to its potential economic and environmental benefits. Compared with traditional uniform spraying, targeted spraying not only reduces chemical costs but also helps protect crops and leaves a significantly smaller environmental footprint in the process. Whole-field weed maps showing the exact locations of weed and crop plants are needed as a "prescription" for spotted herbicide sprayers.

Apart from precision spraying, automatic weed detection has also gained momentum in plant science research. Many agronomy field experiments involve measuring and comparing the density of crops/weeds before and after herbicide application. Traditionally agronomy field experiments are evaluated manually by human observers walking through the field experiment, which is time-consuming, tedious, subjective and prone to rater bias. With the help of automatic weed detection, agronomists can more efficiently and accurately research over larger areas of the fields.

Prior to deep learning, traditional machine learning methods like Support Vector Machines and Linear Discriminant Analysis were used to detect weeds [11][24]. However, extracting features from image pixels can be difficult and requires a substantial amount of domain knowledge in agriculture [8]. Lottes *et al*. [15] and Milioto *et al*. [16] started to experiment with applying deep learning models for semantic segmentation to weed detection and have seen promising success. Convolutional neural networks (CNNs) are a type of deep learning model that is specialized for computer vision tasks such as object detection and semantic segmentation. Object detection identifies objects and marks the location of the objects with bounding boxes. In the case of weed detection, weeds could grow close to the plants which can lead to numerous bounding boxes colliding. Segmentation avoids this problem by performing a pixel-level classification of the image. Semantic segmentation is ideal for weed detection since it not only identifies the weeds but also gives a precise description of their shape and location.

Current applications of CNNs in computer vision are highly task specific. The model generally needs to be trained with a large and well-annotated dataset created for

the desired task. For popular tasks like autonomous driving, training on available public datasets like Cityscapes [5] can often yield satisfactory results. Weed detection datasets have fewer and larger images. Whole-field images are commonly collected by taking a series of images with an unoccupied aerial vehicle (UAV) and stitching these tiles into an whole-field orthomosaic image. With high-resolution cameras, the stitched whole-field orthomosaic image can have a size over 10 GB which makes direct training impractical due to limited GPU VRAM memory sizes.

One practical way to adapt field crop datasets to the training process is to break each image into smaller and uniform-sized tiles [20]. This brings out the problem of balancing the tile size and the amount of overlap between tiles for efficient and effective training. There are a few tiling hyperparameters to consider, mainly size/resolution of image tiles and degree of tile overlap. Batches of the largest tile size that can fit into the GPU memory is preferred for training efficiency [18]. Tile overlaps can potentially increase the dataset size, but have the trade-off of incurring longer training time. Once the training is complete, we would also need a practical way to infer on the much larger whole-field image. One of the challenge is to stitch and merger overlapping tile predictions. Whole-field predictions are particularly challenging for agronomy field experiments which are much more variable than production fields due to imposed variation between regions of the field due to differences in treatment, *e.g*. rate and type of herbicide application.

In this paper, we propose a practical workflow to create an accurate weed map of a wheat field. The workflow describes the end-to-end process of collecting a whole-field image, labelling samples, tiling and training, and inferring on the whole-field image. Six smaller samples are cropped from the whole-field image and manually labelled to train and evaluate the CNN model. The trained model is then applied to the whole-field image to create the desired weed map. There are two canonical and widely-used CNN-based segmentation architectures, UNet and DeepLab, but it is not clear which is most effective for crop-weed segmentation. We also experiment with different tiling overlap strategies and compare the models' performance at different image resolutions. This workflow ensures the model's compatibility with whole-field applications with minimal data collection and annotation effort.

2. Related Works

A fully-convolutional neural network (FCN) [14] is a type of deep learning model that replaces the dense layers in CNNs with convolution layers. FCNs expand the feature maps from the encoder to generate high-resolution pixel-wise predictions for semantic segmentation tasks. The UNet [19] is a popular variation of FCN due to its structural simplicity and effectiveness. The DeepLab [3] model family uses dilated convolution to create deeper and more powerful models and the latest DeepLabv3+ [4] iteration achieved state-of-the-art results on the PASCAL VOC 2012 and Cityscapes datasets.

The success of deep learning in computer vision tasks inspired researchers to apply these techniques to weed detection. Many explored object detection methods like Sapkota *et al*. [21] and Gao *et al*. [7] which works well for sparse vegetation. As vegetation coverage increases, the increased effort in labelling and decreased interpretability due to dense and colliding bounding boxes make object detection less practical for many crop fields. Some publicly available datasets for weed object detection include RoboWeedSupport [6], Rumex-Ancenis Dataset [13], and Maize Seedling and Weeds [17].

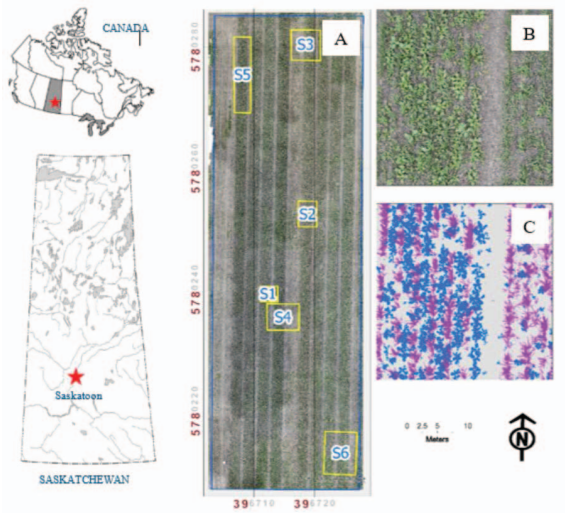


Figure 1: The whole-field orthomosaic RGB image of the wheat field experiment used in this study (A) with sub-region images denoted S1-S6 (outlined with yellow boxes). A cropped and zoomed-in view of one sub-region showing the RGB image (B) and corresponding annotations (C) for wheat crop (purple), weed (blue), and bare soil (grey).

Most weed detection applications, *e.g*. selective spraying, do not need to differentiate individual plant instances, therefore semantic segmentation is an appropriate approach for weed mapping. One popular public dataset for weed semantic segmentation is the Bonn 2016 Sugar Beets Dataset [2]. This dataset uses a ground vehicle to capture close-up images on a sugar beet farm with a four-channel multispectral camera. Lottes *et al*. [15] discussed a sequential FCN-based system trained on the Bonn 2016 Dataset, which generalizes well on two other private datasets. Wang *et al*. [23] found that the inclusion of NIR channels under weak lighting conditions improves segmentation results. You *et al*. [25] suggested that adding hybrid di-

lated convolution and DropBlock to the encoder, and utilizing RGB+NIR based indices, bridge attention blocks and spatial pyramid refinement block can all attribute to increased accuracy. Other notable weed semantic segmentation datasets include the GrassClover Dataset [22] and the WeedMap Dataset [20]. Both the Bonn Sugar Beets and the GrassClover are collected with ground roaming cameras where each image in the dataset covers a relatively small area of the field. The WeedMap dataset is collected with a UAV and provides data in both orthomosaics and tile images. This dataset has sparse vegetation coverage similar to the Bonn Sugar Beets dataset and relatively low resolution. We did not find an existing weed semantic segmentation dataset that is high-resolution and has high weed density and high vegetation coverage from more mature crop growth stages. Most of the research above also only focus on common computer vision evaluation metrics, *e.g*. mIoU, which do not scale well with larger fields due to annotation cost. Therefore, our dataset and the practical and scalable evaluation method for field-scale weed maps is a contribution of this paper.

When dealing with high-resolution images, it is common to cut the large image into tiles due to limited GPU VRAM memory capacity. Reina *et al*. [18] found tiling hyperparameters including size, overlap and orientation can all cause variations in predictions for medical images. Huang *et al*. [9] compared three strategies for stitching tile predictions: clipping, averaging and concatenation and suggest clipping the edge of the tiles for remote sensing tasks. In this study, we will also explore the effect of tiling hyperparameters and image resolution on model training.

3. Methods

3.1. Field Description

The field trial used for this study covered an area of 1503.56 m^2 and was located at the Kernen Crop Research Farm, SK at the Nasser Site (52°16' N, 106°55' W), Saskatoon, SK (Figure 1).

The field trial was seeded on June 29, 2020 with wheat (Triticum aestivum L.) at a rate of 75 seeds per m^2 on a 30.5 cm row spacing. The weed species were seeded in between the crop rows and extra spacing was inserted between individual weed species to prevent overlap for image acquisition. Kochia (Bassia scoparia (L.) A.J. Scott), wild oat (Avena fatua L.), wild mustard (Sinapis arvensis L.), and false cleavers (Galium spurium L.) were cross-seeded in 2 m strips across the experimental area in a split-block design (2.25×2.25 m per plot). The weed species were seeded at the following rates: kochia at 15 kg/ha, wild oat at 90 kg/ha, wild mustard at 8 kg/ha, and false cleavers at 8 kg/ha.

Herbicide treatments were applied perpendicular to the weed strips in a randomized block design (RCBD) with four replicates. This totalled 15 herbicide treatments and 60 plots per weed species. A CO_2 propelled back-pack sprayer was used to apply the herbicide treatments on June 27, 2020, which was calibrated to apply a carrier volume of 100 l ha^{-1} at 40 psi at 6 km/hr.

3.2. UAV Data Acquisition

A total of 215 image frames were acquired using an unoccupied aerial vehicle (UAV) DJI M600 hexacopter UAV (SZ DJI Technology Co., Ltd, Shenzhen, China) equipped with a phase one IXU 1000 RGB camera (Phase One, Copenhagen, Denmark). The images were collected on July 23, 2020 (24 days after seeding) from an altitude of 30 m above ground level with a nadir view while maintaining 80% frontal and 80% side image overlap. At this time, wheat was in the vegetative stage. Detailed information on the imagery is presented in Figure 2.

Weed species were seeded in strips but many volunteer weeds were also in crop rows. As can be seen in Figure 1C, weeds were infesting the crop and distributed evenly over the plots.

Image pre-processing involves mosaicking and masking the region of interest. Image mosaicking was conducted in Pix4D mapper Pro (Pix4D SA, 1015 Lausanne, Switzerland). This step includes image matching, bundle block adjustment, radiometric calibration and orthomosaic generation (in tiff format). The image clipping tool on ArcGIS Pro was then used to mask regions outside the field boundary.

Image characteristics	
Frame size	1108x8708 pixels
Pixel size	2 x 2 cm
Bit depth	24
Bands	3 (Red, Green, Blue)

Figure 2: The UAV DJI M600 hexacopter platform used for image collection, equipped with a phase one IXU 1000 RGB camera and some characteristics of the image.

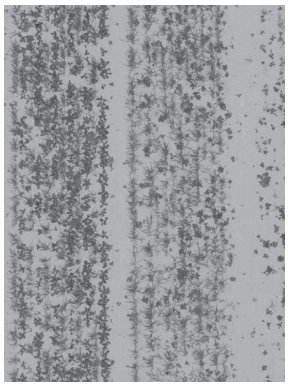
3.3. Data Annotation

Labels on crop, weed, and ground were collected on six different random sub-regions (141 m^2) accounting for 9.37% of the total field area (Figure 1A). An example of the labelled sub-region is presented in Figures 1B and 1C. From the random sub-regions, the foreground vegetation and background soil are separated by thresholding on a vegetation index called Color Index of Vegetation (CIVE) [12]. CIVE is a vegetation index that was based on the principal component analysis from RGB bands and calculated as:

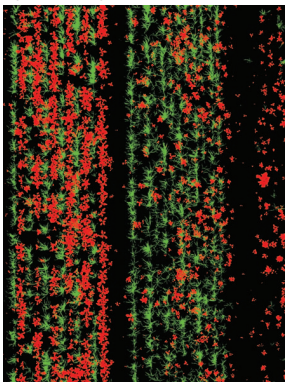
$$CIVE = 0.441 * Red - 0.811 * Green + 0.385 * Blue + 18.78745 \quad (1)$$

From the vegetation class, crops and weeds were separated manually. The manual annotation was conducted by agronomist on vegetation regions to split between weed and crop pixels. Overall, there are six labelling sub-regions. The labelling step was conducted on ArcGIS Pro (version 2.5.0). An example of the labels is presented in Figure 1C. A sub-region RGB image, its CIVE index and the ground-truth annotations are shown in Figure 3.

(a) RGB sub-region image



(b) CIVE transform



(c) Ground-truth annotations

Figure 3: An example sub-region image shown in RGB (a), in the CIVE transform visualization scaled to [0, 255] (b), and with ground-truth annotations (c) for wheat crop (green), weed (red), and bare soil (black).

It is not practical to manually create pixel-level labelling for the entire whole-field orthomosaic image due to its size. The alternative solution is to randomly generate a number of points that are uniformly distributed in the field. These points are manually labelled by plant scientists which represents an independent and out-of-distribution test from the above mentioned test sub-region images annotated by CIVE thresholding. Among the 10,000 randomly generated points, 8,140 points are background pixels, 1,343 are crop pixels and the remaining 517 points are weed pixels. An example of the point labels layered on top of the ground-truth annotation is shown in Figure 4.

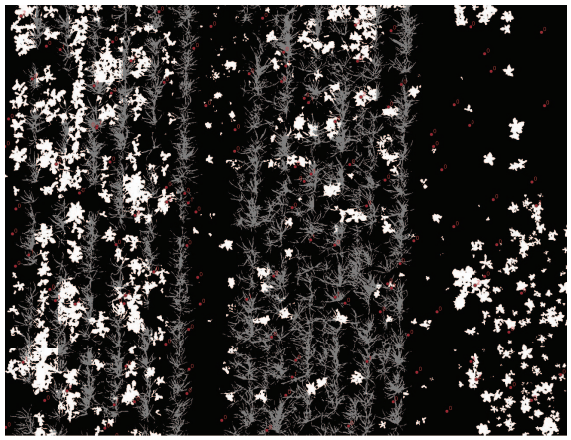


Figure 4: A cropped portion of the whole-field image showing the weed map prediction with background in black, crop in grey, and weed in white. Red dots denote points that were sampled and manually annotated for whole-field image evaluation.

3.4. Weed Detection Model

3.4.1. Sub-region Images Training and Evaluation

The RGB whole-field image used in this project has a dimension of 29,395×90,599 covering the whole controlled field. Six smaller sub-region images were cut out from the original image (Figure 1A) and each pixel was labelled with one of the three classes: background, crop and weed. Three images were chosen for training while one is used for validation and two are used for testing. Even though the dataset consists of only six images, the images cropped from the original ultra-high resolution whole-field image contain more than nine million pixels, which means they can be further cropped into many smaller tiles with good details. These tiles will provide the models with sufficient data to learn the traits of the field.

The sub-region images from the dataset are still too large to fit in most GPUs for training and as a result, are cropped into smaller tiles. The process of training and evaluation is broken into three main parts. First, the train, validation and test images and their labelled masks are broken down into uniform-sized tiles, forming the train, val and test sets. Then, the model is trained on the tiled train and val sets and then used to predict the test tiles. The last step involves stitching predicted test set tiles back into the same size as the test sub-region images. The stitched predictions are

compared to the test images' labels to evaluate the model's performance by calculating class IoUs and mIoU.

There are two strategies that can be used for tiling: divide the images equally on each side or cut fixed-sized tiles from the images. Since all of the images in the dataset have different sizes, the equally dividing method would need the images to be scaled and padded to the same size first. Scaling images will cause loss of information and result in images having different qualities. To avoid losing information from resizing, the images are uniformly cropped into 256×256 tiles. The cropping software is designed such that a tile can be from 0 to less than 100 percent overlapped with the previous tile. An example of three consecutive tiles can be found in Figure 5. To stitch the tiles correctly, the tiles are fed into the model in spatial order and the stitching is done in reverse order using the recorded tile locations. When stitching the predictions with overlapping areas, we employ a greedy approach where the predicted pixels with higher confidence (softmax probability) are kept. When tiling without overlaps, some objects will be split by two or more tiles with each part of the object appearing at the edge of the tiles. The smaller object with an incomplete contour makes it harder for the model to predict. By having overlaps when tiling, an object at the edge of one tile will move to the center of the consecutive one. Another benefit of overlapping tiles is that it creates more data from each sub-region image and expands the training data.

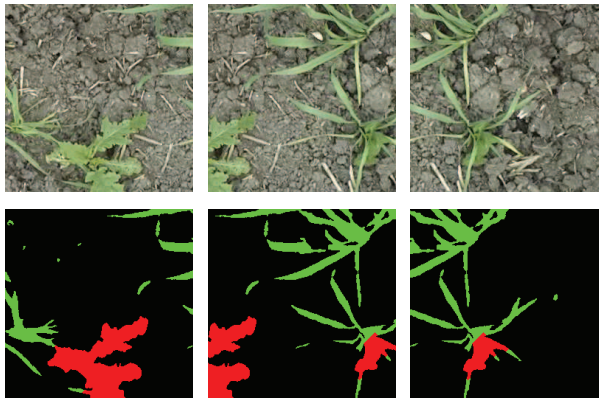


Figure 5: Three consecutive 256×256 tiles with 50% horizontal overlap.

The models used to detect weeds are fully-convolutional neural networks (FCNs) which specialize in semantic segmentation. Instead of a last fully-connected layer which produces a single prediction, the FCN consists of 1×1 convolutions followed by a number of upscaling operations to scale the prediction heatmap back into the same size as the input image. Two popular FCN models are chosen for this project: UNet and DeepLabv3+. The UNet model is implemented with four symmetrical contraction and expansion blocks with Keras, trained with a dynamic learning rate from 1e-4 to 1e-8, batch size of 8 and Adam optimizer. The official implementation and runner scripts of DeepLabv3+ are used with the xception65 backbone. The experiments are run on either a Tesla V100 or an RTX 2080Ti GPU.

3.4.2. Whole-field Evaluation

We use the trained models to predict and create weed maps for the 29,395×90,599 whole-field orthomosaic image. This image goes through a two-stage cropping process where it is first split into 25 smaller images of size 7,348×22,649 and then each smaller image is tiled into 256×256 tiles. The models predict on the tiles and a two-stage stitching process is used to create a weed map of the same size as the whole-field orthomosaic image. This two-stage crop/stitch process allows us to test the pipeline quickly on individual small regions rather than the entire whole-field image, and to easily conduct visual sanity checks with GIS software. The weed-map images are then geo-referenced using the same coordinates from the original whole-field orthomosaic raster. Geo-refencing the predictions gives them practical uses since now every predicted pixel has an exact geographical location which a precision sprayer can use to find weeds. We evaluate the geo-referenced whole-field predictions by extracting the predicted pixel values using the coordinates of the labelled sample points and compare the extracted prediction values with the manually labelled sample points.

3.5. Evaluation Metrics

3.5.1. General Image Segmentation Metrics

There are many metrics available when evaluating the performance of segmentation models. Pixel accuracy is the most simple solution where it gives a percentage of location-wise correct pixels. However, it can be misleading since this method does not consider the relationship between neighbouring pixels when calculating the correctness of a single pixel. For a background-dominated image, a prediction of all backgrounds would have a high pixel accuracy. But this prediction in practice is bad since the model failed to detect any target objects. Other commonly used metrics for classification such as precision and recall all suffer the same problem. Mean Intersection over Union (mIoU) is one of the most commonly used metrics for semantic segmentation. This method is widely used for semantic segmentation as it not only counts for pixel prediction accuracy but also considers spatial accuracy. To calculate class IoU, we use three statistics from the confusion matrix: True Positives (TP), False Positives (FP) and False Negatives (FN). The individual class IoU is given as:

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

Class	UNet	DeepLabv3+
Background	0.990	0.968
Crop	0.935	0.806
Weed	0.694	0.678
Overall	0.873	0.817

Table 1: Comparing the performance of UNet and DeepLabv3+ on the test set with 50% tile overlap. Both models have similar background and weed IoU while UNet has a much higher crop IoU.

The mIoU is calculated as an average over all the classes.

3.5.2. Plant Science-Specific Metrics

In addition to the general image segmentation mIoU metric, we also employ a number of plant science specific evaluation metrics, including user's accuracy, producer's accuracy, and kappa. The above mentioned mIoU metric is not practical when evaluating the full-sized weed map prediction of the entire experimental field. This is due to the significant effort required to manually generate pixel-wise labelling for the large whole-field orthomosaic image. Therefore we adopt a more efficient random point sampling strategy commonly found in GIS software like ArcGIS. A number of randomly sampled points are evenly distributed on the whole-field orthomosaic image and given manual labels. The labelled points are compared with the corresponding model prediction values to generate a confusion matrix. Three types of accuracy indices are calculated to evaluate the prediction quality. The user's accuracy(U_{acc}) is used to indicate false positives for a class while the producer's accuracy(P_{acc}) indicates false negatives. For each class, the user's accuracy and producer's accuracy are given as:

$$U_{acc} = \frac{TP}{TP + FP} \tag{3}$$

$$P_{acc} = \frac{TP}{TP + FN} \tag{4}$$

The kappa index provides an overall assessment of the prediction by comparing the observed accuracy(P_o) with the expected accuracy(P_e). The kappa index is given as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{5}$$

4. Results

4.1. Weed Detection Accuracy

Both the UNet and Deeplabv3+ models are first evaluated on the test sub-region images. Both models' IoU performances when using a tile overlap of 50% are reported in Table 1.

The weed IoU for both models are similar with 0.694 for the UNet and 0.678 for the DeepLabv3+. Both models also showed very high accuracy for background recognition with IoUs above 0.96. However, the DeepLabv3+ resulted in a lower crop IoU of 0.806 which is over 10% lower than the UNet's 0.935. The difference in crop IoU resulted in the UNet having a more than 5% mIoU advantage over the more complex DeepLabv3+.

4.2. Effect of Tile Overlap

The effect of overlapping region percentages on model accuracy is shown in Figure 6. Four overlap percentages are tested: 0%, 25%, 50% and 75%. The mIoU, weed and crop IoUs all increased with larger overlap percentages. This benefit is more visible for the weed class where the UNet has seen a 9% IoU increase. For the classes with higher IoUs, the tiling has less effect on accuracy. Both models had less than 2% IoU gain for crops and no significant increase in background IoU. UNet showed higher IoU in most classes across all overlap percentages except for the weeds. The DeepLabv3+ model showed higher weed IoU with 0% and 25% overlaps while UNet performed better with larger overlap percentages.

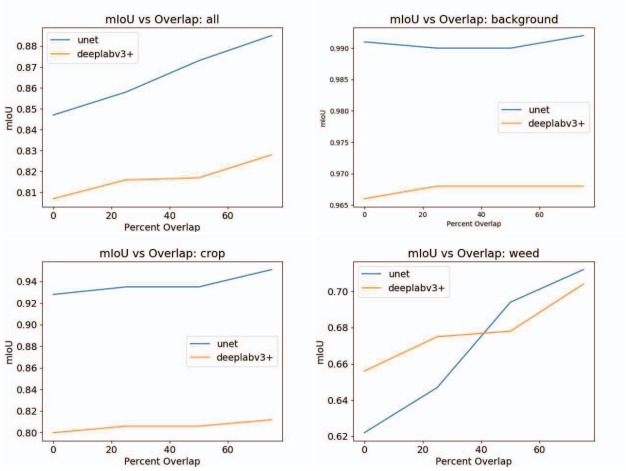


Figure 6: IoU improvements with increased tile overlaps for each class. The background and crop IoU show little to no improvement with more tile overlaps while weed IoU shows positive correlation with tile overlaps.

4.3. Effect of Image Resolution

The flying height of the UAV determines the area the captured image can cover and the quality of the image. Flying at a higher altitude allows the UAV to capture a larger field at the price of a lower-quality image. The effect of capturing at a higher altitude is simulated by downscaling the images. The downscaling experiment is conducted with

Class	Bg	Crop	Weed	Total	U_{acc}
Bg	8089	226	63	8378	0.966
Crop	18	1089	31	1138	0.957
Weed	33	28	423	484	0.874
Total	8140	1343	517	10000	0
P_{acc}	0.994	0.811	0.818	0	0.960

Table 2: Whole-field validation confusion matrix for UNet, $\kappa = 0.867$

Class	Bg	Crop	Weed	Total	U_{acc}
Bg	7967	199	35	8201	0.971
Crop	108	1051	20	1179	0.891
Weed	65	93	462	620	0.745
Total	8140	1343	517	10000	0
P_{acc}	0.979	0.783	0.894	0	0.948

Table 3: Whole-field validation confusion matrix for DeepLabv3+, $\kappa = 0.834$

four levels: original, 2×, 4× and 8×. The results in Figure 7 show both models' performance degrades with higher downscaling. UNet showed less IoU decrease compared to DeepLabv3+ for background and crops. UNet's background accuracy barely degraded while the DeepLabv3+ had an 8% IoU decrease. Both models had more significant degradation in crop IoU with an 8% decrease for UNet and a 39% for DeepLabv3+ at 8× downscaling. Both models showed similar significant degradation in weed IoU with decreases of more than 30% at 8× downscaling. Overall, the UNet appeared to be more robust to downscaling compared with DeepLabv3+ where the latter suffered significantly with more than 2× downscaling.

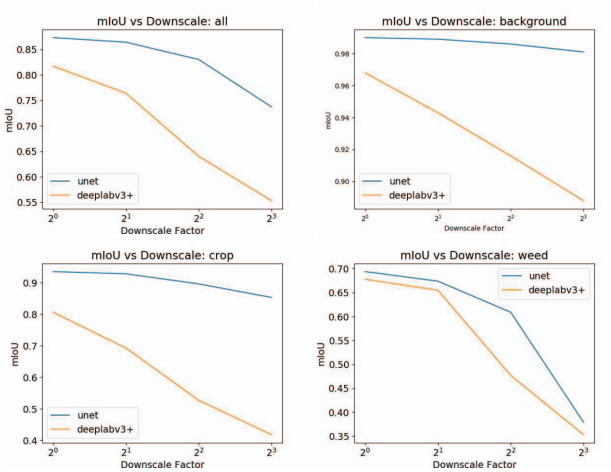


Figure 7: Class IoU degrades with more downsampling. The UNet model shows less degradation in background and crop IoU compared to DeepLabv3+. Both models show visible weed IoU degradation with UNet drops sharply at 4× downsampling and DeepLabv3+ at 2× downsampling.

4.4. Whole-field Prediction

The confusion matrices show the whole-field prediction accuracy of our two models in Table 2 and Table 3. The user's accuracy (U_{acc}), the producer's accuracy (P_{acc}) and the kappa index are calculated for each model.

5. Discussion

5.1. Evaluation Set Results

As shown in section 4.1.1, the UNet model outperforms the more complex DeepLabv3+ for all classes. While both models achieved comparable background and weed IoU, the UNet model has a significantly higher crop IoU than DeepLabv3+. This result is interesting because DeepLabv3+ is a more advanced and complex model which performed better than its predecessors including UNet on large public datasets such as the PASCAL VOC 2012. It should be pointed out that the weed detection tasks have considerably fewer classes compared to the PASCAL dataset which has 20 classes. The weed detection dataset also has another unique trait where background and foreground(vegetation) pixels have distinct colors. This makes the weed detection task an easier task where the main challenge lies in discriminating between the weeds and crops, compared to a generic segmentation task where the model needs to learn the traits of 20 or more distinct objects such as planes and sheep. The reason for the DeepLabv3+ performing worse than the simpler UNet model may be that it is too complex for a simpler task like weed detection and the model overfits the data. Moreover, the smaller size of our dataset, which is a common property of plant image datasets, may also result in the exaggeration of the overfitting problem. UNet on the other hand, being a less capable model, is less likely to memorize the patterns of the smaller dataset resulting in better performance.

The tile overlap result in section 4.1.2 suggests that more tiling overlap results in better performance for both models while UNet shows more improvements compared to DeepLabv3+. The DeepLabv3+ model showed better performance with fewer data samples (0 and 25% overlaps) for weed detection. This is the only time in our experiments that the DeepLabv3+ outperforms the UNet model. It should also be pointed out that the performance boost of higher overlap percentages comes with the cost of more training samples and longer training time. The number of training samples increased from around 700 samples for non-overlapping tiles to almost 3,000 samples for 50%

overlapping tiles. The quadratic increase of samples with overlap percentages could scale poorly for larger datasets. In this experiment, we found that a 50% overlap is a good compromise between training time and model performance.

The downscaling tests in section 4.1.3 show that DeepLabv3+ is more sensitive to lower image quality where its performance degrades more sharply with higher downscaling factors. Both models experience a mild degradation in performance at $2\times$ downscaling and a sharper drop in IoU with more than $4\times$ downscaling. This could be caused by too much spatial information being lost with high downscaling factors. This means that the drone could fly twice as high with a small sacrifice in model performance whereas a flight height over 4 times could substantially impact the prediction accuracy.

5.2. Whole-field Prediction Results

Our sampling results show that both models generalize well on the whole-field orthomosaic image. Both U_{acc} and P_{acc} are above 95% for the background class which shows the model can easily discriminate between background and vegetation pixels. One noticeable difference between the two models is the DeepLab model is more aggressive in predicting pixels as weeds. The DeepLab model predicted 620 samples as weed while the UNet predicted 484 samples as weed. This led to the DeepLab model having fewer false negatives but more false positives on the weed class compared with the UNet. Overall, UNet has a higher kappa index of 0.867 compared with a kappa index of 0.834 for DeepLab. The whole-field accuracy results agree with the mIoU of the evaluation set indicating that the simpler UNet model generalizes better than the more complex DeepLab model on our dataset.

5.3. Study Limitation and Future Work

There are some limitations in this work that we discuss in this section and propose to address as future work. The experimentation mentioned in this paper is done on one single wheat field. Although this method should be adaptable to other fields even with different crop and weed species, we cannot confirm this due to a lack of experimental evidence. We would like to extend our experiments to other farm fields in the future to evaluate the generalizability of our method. Moreover, we have plans to explore the possibility of transferring a model trained from one field to another field. This potential domain adaptation research could take advantage of a pre-trained model and would likely lead to better performance for lower annotation costs.

Another limitation of this work is that the experiments are done in a controlled experimental field which is different from production farms. We have more control of the weed species, the types of herbicide treatments and the density of vegetations in our experimental field. If future projects, we would like to experiment on production farms where the field condition is less carefully managed. In the proposed production farm experiments, we would also like to collect data and study the different growth stages of those fields where vegetation coverage and sparsity could vary greatly.

It is also important to note that our downscaling simulation is only an approximation of the effect of higher altitude UAV flights. Our experiment simply downscales the images. Even though the pixel quality would be similar, flying at higher altitude also results in the image covering more ground. The whole-field image from a higher altitude UAV would include more rows of vegetations which would increase the amount of information the input image provides. The effect of more objects with worse pixel quality on models will be studied in future research.

Lastly the geo-referencing of the weed map is presumed to be accurate since it is using the exact coordinates of the RGB whole-field orthomosaic image. Ground control points were used during UAV data collection to mitigate geo-referencing errors. However, there may exist warping and stitching artifacts from the orthomosaic generation process. Such distortions would cause the weed map's geolocation to not align properly with the actual field. The exact accuracy of geo-referencing an orthomosaic and methods to correct the alignment issue will be explored in future.

6. Conclusion

In this paper, we presented a practical and effective workflow to create a weed map using CNNs on high resolution UAV image of a wheat field. We started by collecting an image of the whole controlled wheat field. Then six sub-region images were cropped from the whole-field orthomosaic image and manually labelled to form our training and testing sets. Two CNN models, the UNet and the DeepLabv3+ were trained and evaluated using the annotated sub-region images. The trained models were then used to create full-sized weed maps for the entire whole-field image and the weed maps were evaluated by comparing them with labels generated by randomly sampling 10,000 points from the whole-field orthomosaic image.

We achieved an 87% mIoU and a 69% weed IoU on the test sub-region images. We also showed that with a sacrifice in longer training time more tile overlaps result in better performance and more than $4\times$ downscaling significantly impacts the model accuracy. Our whole-field evaluation results showed that our model trained on only a few small sub-region images from the whole-field orthomosaic image can generalize very well on the entire whole-field orthomosaic image. Our dataset and software will be made publicly available, and we hope our research will spark more interest in computer vision in agriculture and serve as a guideline for other field weed mapping research.

References

[1] Bhagirath Singh Chauhan. Grand challenges in weed management, 2020.

[2] Nived Chebrolu, Philipp Lottes, Alexander Schaefer, Wera Winterhalter, Wolfram Burgard, and Cyrill Stachniss. Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *The International Journal of Robotics Research*, pages 1045–1052, 2017.

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[6] Mads Dyrmann, Rasmus Nyholm Jørgensen, and Henrik Skov Midtiby. Roboweedsupport-detection of weed locations in leaf occluded cereal crops using a fully convolutional neural network. *Advances in Animal Biosciences*, pages 842–847, 2017.

[7] Junfeng Gao, Andrew P French, Michael P Pound, Yong He, Tony P Pridmore, and Jan G Pieters. Deep convolutional neural networks for image-based convolvulus sepium detection in sugar beet fields. *Plant Methods*, pages 1–12, 2020.

[8] ASM Mahmudul Hasan, Ferdous Sohel, Dean Diepeveen, Hamid Laga, and Michael GK Jones. A survey of deep learning techniques for weed detection from images. *Computers and Electronics in Agriculture*, page 106067, 2021.

[9] Bohao Huang, Daniel Reichman, Leslie M Collins, Kyle Bradbury, and Jordan M Malof. Tiling and stitching segmentation output for remote sensing: Basic challenges and recommendations. *arXiv preprint arXiv:1805.12219*, 2018.

[10] Daniel Kane and LLC Solutions. Carbon sequestration potential on agricultural lands: a review of current science and available practices. *National Sustainable Agriculture Coalition Breakthrough Strategies and Solutions, LLC*, pages 1–35, 2015.

[11] Y Karimi, SO Prasher, RM Patel, and SH Kim. Application of support vector machine technology for weed and nitrogen stress detection in corn. *Computers and electronics in agriculture*, pages 99–109, 2006.

[12] Takashi Kataoka, Toshihiro Kaneko, Hiroshi Okamoto, and S Hata. Crop growth estimation system using machine vision. In *Proceedings IEEE/ASME international conference on advanced intelligent mechatronics (AIM 2003)*, pages b1079–b1083, 2003.

[13] Tsampikos Kounalakis, Georgios A Triantafyllidis, and Lazaros Nalpantidis. Deep learning-based visual recognition of rumex for robotic precision farming. *Computers and Electronics in Agriculture*, page 104973, 2019.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[15] Philipp Lottes, Jens Behley, Andres Milioto, and Cyrill Stachniss. Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. *IEEE Robotics and Automation Letters*, pages 2870–2877, 2018.

[16] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In *IEEE international conference on robotics and automation (ICRA)*, pages 2229–2235, 2018.

[17] Longzhe Quan, Huaiqu Feng, Yingjie Lv, Qi Wang, Chuanbin Zhang, Jingguo Liu, and Zongyang Yuan. Maize seedling detection under different growth stages and complex field environments based on an improved faster r–cnn. *Biosystems Engineering*, pages 1–23, 2019.

[18] G Anthony Reina, Ravi Panchumarthy, Siddhesh Pravin Thakur, Alexei Bastidas, and Spyridon Bakas. Systematic evaluation of image tiling adverse effects on deep learning semantic segmentation. *Frontiers in neuroscience*, page 65, 2020.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceeding of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[20] Inkyu Sa, Marija Popović, Raghav Khanna, Zetao Chen, Philipp Lottes, Frank Liebisch, Juan Nieto, Cyrill Stachniss, Achim Walter, and Roland Siegwart. Weedmap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming. *Remote Sensing*, page 1423, 2018.

[21] Bishwa B Sapkota, Chengsong Hu, and Muthukumar V Bagavathiannan. Evaluating cross-applicability of weed detection models across different crops in similar production environments. *Frontiers in Plant Science*, page 837726, 2022.

[22] Soren Skovsen, Mads Dyrmann, Anders K Mortensen, Morten S Laursen, René Gislum, Jorgen Eriksen, Sadaf Farkhani, Henrik Karstoft, and Rasmus N Jorgensen. The grassclover image dataset for semantic and hierarchical species understanding in agriculture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019.

[23] Aichen Wang, Yifei Xu, Xinhua Wei, and Bingbo Cui. Semantic segmentation of crop and weed using an encoder-decoder network and image enhancement method under uncontrolled outdoor illumination. *IEEE Access*, pages 81724–81734, 2020.

[24] Alexander Wendel and James Underwood. Self-supervised weed detection in vegetable crops using ground based hyperspectral imaging. In *IEEE international conference on robotics and automation (ICRA)*, pages 5128–5135, 2016.

[25] Jie You, Wei Liu, and Joonwhoan Lee. A dnn-based semantic segmentation for detecting weed and crop. *Computers and Electronics in Agriculture*, page 105750, 2020.